

Characterizations of interpretability in bounded arithmetic

Joost J. Joosten
University of Barcelona

February 2, 2016

Abstract

This paper deals with three tools to compare proof-theoretic strength of formal arithmetical theories: interpretability, Π_1^0 -conservativity and proving restricted consistency. It is well known that under certain conditions these three notions are equivalent and this equivalence is often referred to as the Orey-Hájek characterization of interpretability.

In this paper we look with detail at the Orey-Hájek characterization and study what conditions are needed and in what meta-theory the characterizations can be formalized.

1 Introduction

Interpretations are everywhere used in mathematics and mathematical logic. Basically, a theory U interprets a theory V –we write $U \triangleright V$ – whenever there is some translation from the symbols of the language of V to formulas of the language of U so that under a natural extension of this translation the axioms of V are mapped to theorems of U .

The corresponding intuition should be that U is at least as strong or expressible as V . And indeed, interpretations are used for example to give relative consistency proofs or to establish undecidability of theories. As such, interpretations are considered an important metamathematical notion. Probably, the first time that interpretations received a formal and systematic treatment has been in the book by A. Tarski, A. Mostowski and R. Robinson ([17]). In the current paper we will study that notion of interpretability and also some related notions. Sometimes we speak of *relative* interpretability as to indicate that quantifications become relativized to some domain specifier as we shall define precisely later on.

We will relate the notion of relative interpretability to two other basic metamathematical notions. The first such notion is the notion of consistency. The notion of consistency is central to mathematical logic and considered key and fundamental.

A second notion is that of Π_1^0 conservativity. Below we will exactly define what Π_1^0 formulas are, but basically, those are formulas in the language of arithmetic which are of the form $\forall x \psi(x)$ where ψ is some decidable predicate. On the other hand, Σ_1^0 formulas

are those of the form $\exists x\psi(x)$ for decidable ψ . Since all true theories prove exactly the same set of Σ_1^0 sentences, the first natural and interesting class of formulas to distinguish theories is on the Π_1^0 level. Therefore, the notion of Π_1^0 conservativity has been very central in mathematical logic and foundational discussions. We say that a theory U is Π_1^0 conservative over V whenever any Π_1^0 sentence provable by V is also provable by U .

The main purpose of this paper is to discuss how these three different notions are related to each other in certain circumstances. This relation is known as the *Orey-Hájek* characterization of relative interpretability.

As such, the paper contains many well-known results and various formulations are taken from [8]. However, we think that it is instructive that all these results are put together and moreover that a clear focus is on the requirements needed so that various implications are formalizable in weak theories.

Apart from the main focus –which is bringing together facts of the Orey-Hájek characterization of relative interpretability and formalizations thereof– the paper contains a collection of new observations that might come in handy. For example, our simple generalization of Pudlák’s lemma as formulated in Lemma 5.5 has been a main tool in proving arithmetical correctness of a new series of interpretability principles in [9].

2 Preliminaries

As mentioned before, a central notion in this paper is that of consistency. Consistency is a notion that concerns syntax: no sequence of symbols that constitute a proof will yield the conclusion that $0 = 1$. It shall be an important criterion whether or not a theory proves the consistency of another. As such we want that theories can talk about syntax.

The standard choice to represent syntax is by Gödel numbering, assigning natural numbers to syntax. Thus, our theories should contain a modicum of arithmetic. In this section we shall make some basic observations on coding and then fix what minimal arithmetic we should have in our base theory. We shall formulate some fundamental properties of this base theory and refer to the literature for further background. Further, we shall fix the notation that is used in the remainder of this paper.

2.1 A short word on coding

Formalization calls for coding of syntax. At some places in this paper we shall need estimates of codes of syntactical objects. Therefore it is good to discuss the nature of the coding process we will employ. However we shall not consider the implementation details of our coding.

We shall code strings over some finite alphabet A with cardinality a . A typical coding protocol could be the following. First we define an alphabetic order on A . Next we enumerate all finite strings over A in the following way (pseudo-lexicographic order).

To start, we enumerate all strings of length 0, then of length 1, etcetera. For every n , we enumerate the strings of length n in alphabetic order. The coding of a finite string over A will just be its ordinal number in this enumeration. We shall now see some easy arithmetical properties of this coding. We shall often refrain from distinguishing syntactical objects and their codes.

1. There are a^n many strings of length n .
2. There are $a^n + a^{n-1} \dots + 1 = \frac{a^{n+1}-1}{a-1}$ many strings of length $\leq n$.
3. From (2) it follows that the code of a syntactical object of length n , is $\mathcal{O}(\frac{a^{n+1}-1}{a-1}) = \mathcal{O}(a^n)$ big.
4. Conversely, the length of a syntactical object that has code φ is $\mathcal{O}(|\varphi|)$ (logarithm/length of φ) big.
5. If φ and ψ are codes of syntactical objects, the concatenation $\varphi \star \psi$ of φ and ψ is $\mathcal{O}(\varphi \cdot \psi)$ big. For, $|\varphi \star \psi| = |\varphi| + |\psi|$, whence by (3), $\varphi \star \psi \approx a^{|\varphi|+|\psi|} = a^{|\varphi|} \cdot a^{|\psi|} = \varphi \cdot \psi$.
6. If φ and t are (codes of) syntactical objects, then $\varphi_x(t)$ is $\mathcal{O}(\varphi^{|t|})$ big. Here $\varphi_x(t)$ denotes the syntactical object that results from φ by replacing every (unbounded) occurrence of x by t . The length of φ is about $|\varphi|$. In the worst case, these are all x -symbols. In this case, the length of $\varphi_x(t)$ is $|\varphi| \cdot |t|$ and thus $\varphi_x(t)$ is $\mathcal{O}(a^{|\varphi| \cdot |t|}) = \mathcal{O}(t^{|\varphi|}) = \mathcal{O}(\varphi^{|t|}) = \mathcal{O}(2^{|\varphi| \cdot |t|})$ big.

As mentioned, we shall refrain from the technical characteristics of our coding and refer to the literature for examples. Rather, we shall keep in mind restrictions on the sizes and bounds as mentioned above. Also, we shall assume that we work with a natural poly-time coding with poly-time decoding functions so that the code of substrings is always smaller than the code of the entire string.

2.2 Arithmetical theories

Since substitution is key to manipulating syntax we need, by our observations above, a function whose growth-rate can capture substitution. Thus, we choose to work with the smash function \sharp defined by $x \sharp y := 2^{|x| \cdot |y|}$ where $|x| := \lceil \log_2(x+1) \rceil$ is the length of the number x in binary. We shall often also employ the function ω_1 which is of similar growth-rate and defined by $\omega_1(x) := 2^{|x|^2}$.

Next, we need a certain amount of induction. For a formula φ , the regular induction formula I_φ is given by

$$\varphi(0) \wedge \forall x (\varphi(x) \rightarrow \varphi(x+1)) \rightarrow \forall x \varphi(x).$$

However, it turns out that we can work with a weaker version of induction called polynomial induction denoted by PIND:

$$\varphi(0) \wedge \forall x (\varphi(\lfloor \frac{1}{2}x \rfloor) \rightarrow \varphi(x)) \rightarrow \forall x \varphi(x)$$

or equivalently

$$\varphi(0) \wedge \forall x (\varphi(x) \rightarrow \varphi(2x)) \wedge \forall x (\varphi(x) \rightarrow \varphi(2x+1)) \rightarrow \forall x \varphi(x).$$

The idea is that one can conclude $\varphi(x)$ by only logarithmically many calls upon the induction hypothesis with this PIND principle. For example to conclude $\varphi(18)$ we'd go $\varphi(0) \rightarrow \varphi(1) \rightarrow \varphi(2) \rightarrow \varphi(4) \rightarrow \varphi(9) \rightarrow \varphi(18)$.

Typically, induction on syntax is of this nature and in order to conclude a property of (the Gödel number of) some formula ψ we need to apply the induction hypothesis to the number of subformulas of ψ which is linear in the length of ψ . Thus, most inductions over syntax can be established by PIND rather than the regular induction schema.

Moreover, we shall restrict the formulas on which we allow ourselves to apply PIND to so to end up with a weak base theory. As we shall see, most of our arguments can be formalized within Buss' theory¹ S_2^1 .

The theory S_2^1 is formulated in the language of arithmetic $\{0, S, +, \cdot, \#, |x|, \lfloor \frac{1}{2}x \rfloor, \leq\}$. Apart from some basic axioms that define the symbols in the language, S_2^1 is axiomatized by PIND induction for Σ_1^b formulas. The Σ_1^b formulas are those formed from atomic formulas via the boolean operators, sharply bounded quantification and bounded existential quantification. Sharply bounded quantification is quantification of the form $\mathcal{Q} x < |t|$ for $\mathcal{Q} \in \{\forall, \exists\}$. Bounded existential quantification in contrast, is of the form $\exists x < t$. We refer the reader for [1] or [6] for further details and for the definitions of the related Σ_n^b and Π_n^b hierarchies.

Equivalent to the PIND principle (see [11, Lemma 5.2.5]) is the length induction principle LIND:

$$\varphi(0) \wedge \forall x (\varphi(x) \rightarrow \varphi(x+1)) \rightarrow \forall x \varphi(|x|).$$

So, from the progressiveness of φ , we can conclude $\varphi(x)$ for any x for which the exponentiation is defined. We shall later see that if we are working with definable cuts (definable initial segments of the natural numbers with some natural closure properties) we can without loss of generality assume that exponentiation is defined for elements of this cut.

Although most of our reasoning can be performed in S_2^1 , we sometimes mention stronger theories. As always Peano Arithmetic (PA) contains open axioms that define the symbols $0, S, +$ and \cdot and induction axioms I_φ for any arithmetical formula φ . Similarly, IS_n is as PA where instead we only have induction axioms I_φ for $\varphi \in \Sigma_n$. Here, Σ_n refers to the

¹As mentioned, the substitution operation on codes of syntactical objects asks for a function of growth rate $x^{|x|}$. In Buss's S_2^1 this is the smash function $\#$. In the theory $I\Delta_0 + \Omega_1$ this is the function $\omega_1(x)$. However, contrary to S_2^1 , the theory $I\Delta_0 + \Omega_1$ –aka S_2^- – is not known to be finitely axiomatizable.

usual arithmetical hierarchy (see e.g. [6]) in that such formulas are written as a decidable formula preceded by a string of n alternating quantifiers with an existential quantifier up front. In case no free variables are allowed in the induction formulas, we flag this by a superscript “ $-$ ” as in IS_n^- .

Another important arithmetical principle that we will encounter frequently is collection. For example $\text{B}\Sigma_n$ is the so-called collection scheme for Σ_n -formulae. Roughly, $\text{B}\Sigma_n$ says that the range of a Σ_n -definable function on a finite interval is again finite. A mathematical formulation is $\forall x \leq u \exists y \sigma(x, y) \rightarrow \exists z \forall x \leq u \exists y \leq z \sigma(x, y)$ where $\sigma(x, y) \in \Sigma_n$ may contain other variables too.

The least number principle LF for a class of formulas is the collection $\exists x \varphi(x) \rightarrow \exists x (\varphi(x) \wedge \forall y < x \neg \varphi(y))$ for $\varphi \in \Gamma$.

2.3 Numberized theories

The notion of interpretability applies to any pair of theories and not necessarily need they contain any arithmetic. However, in this paper we will prove that $U \triangleright V$ can in various occasions be equivalent to other properties that are stated in terms of numbers. For example, in certain situations we have that $U \triangleright V$ is equivalent to U proving all the Π_1^0 formulas that V does. Clearly, in this situation we should understand that U and V come with a natural interpretation of numbers.

Definition 2.1. We will call a pair $\langle U, k \rangle$ a *numberized theory* if $k : U \triangleright \text{S}_2^1$. A theory U is *numberizable* or *arithmetical* if for some j , $\langle U, j \rangle$ is a numberized theory.

From now on, we shall only consider numberizable or numberized theories. Often however, we will fix a numberization j and reason about the theory $\langle U, j \rangle$ as if it were formulated in the language of arithmetic.

A disadvantage of doing so is clearly that our statements may be somehow misleading; when we think of, e.g., ZFC we do not like to think of it as coming with a fixed numberization. However, for the kind of characterizations treated in this paper, it is really needed to have numbers around. We shall most of the times work with sequential theories. Basically, sequentiality means that any finite sequence of objects can be coded.

2.4 Metamathematics in numberized theories

On many occasions, we want to represent numbers by terms (numerals) and then consider the code of that term. It is not a good idea to represent a number n by

$$\overbrace{S \dots S}^{n \text{ times}} 0.$$

For, the length of this object is $n + 1$ whence its code is about 2^{n+1} and we would like to avoid the use of exponentiation. In the setting of weaker arithmetics it is common practice

to use so-called *efficient numerals*. These numerals are defined by recursion as follows. $\overline{0} = 0$; $\overline{2 \cdot n} = (SS0) \cdot \overline{n}$ and $\overline{2 \cdot n + 1} = S((SS0) \cdot \overline{n})$. Clearly, these numerals implement the system of dyadic notation which perfectly ties up with the PIND principle. Often we shall refrain from distinguishing n from its numeral \overline{n} or even the Gödel number $\ulcorner \overline{n} \urcorner$ of its numeral.

As we want to do arithmetization of syntax, our theories should be coded in a simple way. We will assume that all our theories U have an axiom set that is decidable in polynomial time. That is, there is some formula $\text{Axiom}_U(x)$ which is Δ_1^b (both the formula and its negation are provably equivalent to a Σ_1^b formula) in S_2^1 , with

$$S_2^1 \vdash \text{Axiom}_U(\varphi) \text{ iff } \varphi \text{ is an axiom of } U.$$

The choice of Δ_1^b -axiomatizations is also motivated by Lemma 2.2 below. Most natural theories like ZFC or PA indeed have Δ_1^b -axiomatizations. Moreover, by a sharpening of Craig's trick, any recursive theory is deductively equivalent to one with a Δ_1^b -axiomatization.

We shall employ the standard techniques and concepts necessary for the arithmetization of syntax. Thus, we shall work with provability predicates \Box_U corresponding uniformly to arithmetical theories U . We shall adhere to the standard dot notation so that, for example, $\Box_U \varphi(\dot{x})$ denotes a formula with one free variable x so that for each value of x , $\Box_U \varphi(\dot{x})$ is provably equivalent to $\Box_U \varphi(\overline{x})$.

We shall always write the formalized version of a concept in sans-serif style. For example, $\text{Proof}_U(p, \varphi)$ stands for the formalization of “ p is a U -proof of φ ”, $\text{Con}(U)$ stands for the formalization of “ U is a consistent theory” and so forth. It is known that for theories U with a poly-time axiom set, the formula $\text{Proof}_U(p, \varphi)$ can be taken to be in Δ_1^b being a poly-time decidable predicate. Again, [1] and [6] are adequate references.

For already really weak theories T we have Σ_1 -completeness in the sense that T proves any true Σ_1 sentence. However, proofs of Σ_1 -sentences σ are multi-exponentially big, that is, 2_n^σ for some n depending on σ . (See e.g., [6].) As such, we cannot expect that we can formalize the Σ_1 completeness theorem in theories where exponentiation is not necessarily total.

However, for $\exists \Sigma_1^b$ -formulas we do have a completeness theorem (see [1]) in bounded arithmetic. From now on, we shall often write a sup-index to a quantifier to specify the domain of quantification.

Lemma 2.2. *If $\alpha(x) \in \exists \Sigma_1^b$, then there is some standard natural number n such that*

$$S_2^1 \vdash \forall x [\alpha(x) \rightarrow \exists p < \omega_1^n(x) \text{Proof}_U(p, \alpha(\dot{x}))].$$

This holds for any reasonable arithmetical theory U . Moreover, we have also a formalized version of this statement.

$$S_2^1 \vdash \forall^{\exists \Sigma_1^b} \alpha \exists n \Box_{S_2^1} (\forall x [\dot{\alpha}(x) \rightarrow \exists p < \omega_1^n(x) \text{Proof}_U(p, \dot{\alpha}(\dot{x}))]).$$

2.5 Consistency and reflexive theories

Since Gödel's second incompleteness theorem, we know that no recursive theory that is consistent can prove its own consistency. For a large class of natural theories we do have a good approximation of proving consistency though. A theory is *reflexive* if it proves the consistency of all of its finite subtheories. Reflexivity is a natural notion and most natural non-finitely axiomatized theories are reflexive like, for example, primitive recursive arithmetic and PA.

Many meta-mathematical statements involve the notion of reflexivity. There exist various ways in which reflexivity can be formalized, and throughout the literature we can find many different formalizations. For stronger theories, all these formalizations coincide. But for weaker theories, the differences are essential. We give some formalizations of reflexivity.

1. $\forall n \ U \vdash \text{Con}(U[n])$ where $U[n]$ denotes the conjunction of the first n axioms of U .
2. $\forall n \ U \vdash \text{Con}(U \upharpoonright n)$ where $\text{Con}(U \upharpoonright n)$ denotes that there is no proof of falsity using only axioms of U with Gödel numbers $\leq n$.
3. $\forall n \ U \vdash \text{Con}_n(U)$ where $\text{Con}_n(U)$ denotes that there is no proof of falsity with a proof p where p has the following properties. All non-logical axioms of U that occur in p have Gödel numbers $\leq n$. All formulas φ that occur in p have a logical complexity $\rho(\varphi) \leq n$.

Here ρ is some complexity measure that basically counts the number of quantifier alternations in φ . Important features of this ρ are that for every n , there are truth predicates for formulas with complexity n . Moreover, the ρ -measure of a formula should be more or less (modulo some poly-time difference, see Remark 3.4) preserved under translations. An example of such a ρ is given in [19].

It is clear that (2) \Rightarrow (3) can be proven in any weak base theory. For the corresponding provability notions, the implication reverses. In this paper, our notion of reflexivity shall be the third one.

We shall write $\Box_{U,n}\varphi$ for $\neg\text{Con}_n(U + \neg\varphi)$ or, equivalently, $\exists p \ \text{Proof}_{U,n}(p, \varphi)$. Here, $\text{Proof}_{U,n}(p, \varphi)$ denotes that p is a U -proof of φ with all axioms in p are $\leq n$ and for all formulas ψ that occur in p , we have $\rho(\psi) \leq n$.

Remark 2.3. An inspection of the proof of provable Σ_1 -completeness (Lemma 2.2) gives us some more information. The proof p that witnesses the provability in U of some $\exists\Sigma_1^b$ -sentence α , can easily be taken so that all axioms occurring in p are about as big and complex as α . Thus, from α , we get for some n (depending linearly on α) that $\text{Proof}_{U,n}(p, \alpha)$.

If we wish to emphasize the fact that our theories are not necessarily in the language of arithmetic, but just can be numberized, our formulations of reflexivity should be slightly changed. For example, (3) will for some $\langle U, j \rangle$ look like $j : U \triangleright S_2^1 + \{\text{Con}_n(U) \mid n \in \omega\}$.

If U is a reflexive theory, we do not necessarily have any reflection principles. That is, we do not have $U \vdash \Box_V \varphi \rightarrow \varphi$ for some natural $V \subset U$ and for some natural class of formulae φ . We do have, however, a weak form of $\forall\Pi_1^b$ -reflection. This is expressed in the following lemma.

Lemma 2.4. *Let U be a reflexive theory. Then*

$$S_2^1 \vdash \forall^{\forall\Pi_1^b} \pi \forall n \Box_U \forall x (\Box_{U,n} \pi(\dot{x}) \rightarrow \pi(x)).$$

Proof. Reason in S_2^1 and fix π and n . Let m be such that we have (see Lemma 2.2 and Remark 2.3)

$$\Box_U \forall x (\neg \pi(x) \rightarrow \Box_{U,m} \neg \pi(\dot{x})).$$

Furthermore, let $k := \max\{n, m\}$. Now, reason in U , fix some x and assume $\Box_{U,n} \pi(x)$. Thus, clearly also $\Box_{U,k} \pi(x)$. If now $\neg \pi(x)$, then also $\Box_{U,k} \neg \pi(x)$, whence $\Box_{U,k} \perp$. This contradicts the reflexivity, whence $\pi(x)$. As x was arbitrary we get $\forall x (\Box_{U,n} \pi(x) \rightarrow \pi(x))$. \dashv

We note that this lemma also holds for the other notions of restricted provability we introduced in this subsection.

3 Formalized interpretability

As we already mentioned, our notion of interpretability is the one studied by Tarski et al in [17]. In that notion, any axiom needs to be provable after translation. Under some fairly weak conditions this implies that also theorems are translated to theorems. However, in the domain of bounded arithmetics we do not generally have this. In the realm of formalized interpretation therefore, there has been a tendency to consider a small adaptation of the original notion Tarski. This adaptation as introduced by Visser is called *smooth* interpretability. In this subsection we shall exactly define this notion and see how it relates to other notions of formalized interpretability. In various ways, one can hold that theorems interpretability as discussed below is actually the more natural formalized version of interpretability.

The theories that we study in this paper are theories formulated in first order predicate logic. All theories have a finite signature that contains identity. For simplicity we shall assume that all our theories are formulated in a purely relational way. Here is the formal definition of a relative interpretation.

Definition 3.1. A *translation* k of the language of a theory S into the language of a theory T is a pair $\langle \delta, F \rangle$ for which the following holds.

The first component δ , is called the *domain specifier* and is a formula in the language of T with a single free variable. This formula is used to specify the domain of our interpretation.

The second component, F , is a finite map that sends relation symbols R (including identity) from the language of S , to formulas $F(R)$ in the language of T . We demand for all R that the number of free variables of $F(R)$ equals the arity of R .² Recursively we define the translation φ^k of a formula φ in the language of S as follows.

- $(R(\vec{x}))^k = F(R)(\vec{x})$;
- $(\varphi \wedge \psi)^k = \varphi^k \wedge \psi^k$ and likewise for other boolean connectives;
(in particular, this implies $\perp^k = \perp$);
- $(\forall x \varphi(x))^k = \forall x (\delta(x) \rightarrow \varphi^k)$ and analogously for the existential quantifier.

A relative interpretation k of a theory S into a theory T is a translation $\langle \delta, F \rangle$ so that $T \vdash \varphi^k$ for all axioms φ of S .

To formalize insights about interpretability in weak meta-theories like S_2^1 we need to be very careful. Definitions of interpretability that are unproblematically equivalent in a strong theory like, say, IS_1 diverge in weak theories. As we shall see, the major source of problems is the absence of $B\Sigma_1$.

In this subsection, we study various divergent definitions of interpretability. We start by making an elementary observation on interpretations. Basically, the next definition and lemma say that translations transform proofs into translated proofs.

Definition 3.2. Let k be a translation. By recursion on a proof p in natural deduction we define the translation of p under k , we write p^k . For this purpose, we first define $k(\varphi)$ for formulae φ to be³ $\bigwedge_{x_i \in \text{FV}(\varphi)} \delta(x_i) \rightarrow \varphi^k$. Here $\text{FV}(\varphi)$ denotes the set of free variables of φ . Clearly, this set cannot contain more than $|\varphi|$ elements, whence $k(\varphi)$ will not be too big. Obviously, for sentences φ , we have $k(\varphi) = \varphi^k$.

If p is just a single assumption φ , then p^k is $k(\varphi)$. The translation of the proof constructions are defined precisely in such a way that we can prove Lemma 3.3 below. For example, the translation of

$$\frac{\varphi \quad \psi}{\varphi \wedge \psi}$$

will be

$$\frac{\frac{\frac{[\bigwedge_{x_i \in \text{FV}(\varphi \wedge \psi)} \delta(x_i)]_1}{\bigwedge_{x_i \in \text{FV}(\varphi)} \delta(x_i)} \quad \frac{\bigwedge_{x_i \in \text{FV}(\varphi)} \delta(x_i) \rightarrow \varphi^k}{\varphi^k} \quad \frac{\mathcal{D}}{\psi^k}}{\varphi^k \wedge \psi^k} \rightarrow I, 1$$

²Formally, we should be more precise and specify our variables.

³To be really precise we should say that, for example, we let smaller x_i come first in $\bigwedge_{x_i \in \text{FV}(\varphi)} \delta(x_i)$.

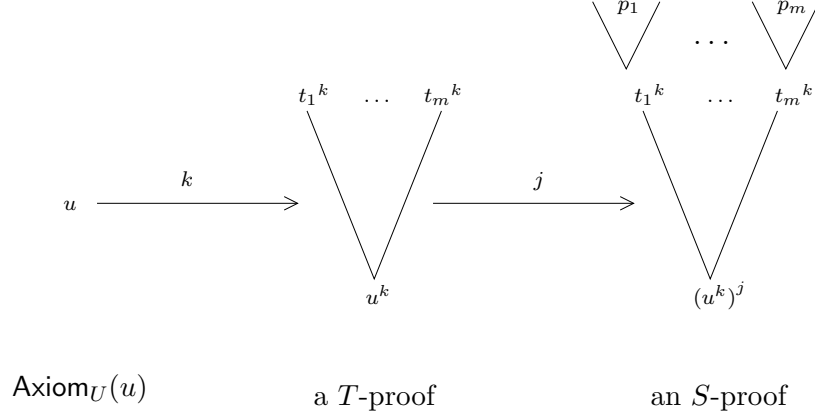


Figure 1: Transitivity of interpretability

where \mathcal{D} is just a symmetric copy of the part above φ^k . We note that the translation of the proof constructions is available⁴ in \mathbf{S}_2^1 , as the number of free variables in $\varphi \wedge \psi$ is bounded by $|\varphi \wedge \psi|$.

Lemma 3.3. *If p is a proof of a sentence φ with assumptions in some set of sentences Γ , then for any translation k , p^k is a proof of φ^k with assumptions in Γ^k .*

Proof. Note that the restriction on sentences is needed. For example

$$\frac{\forall x \varphi(x) \quad \forall x (\varphi(x) \rightarrow \psi(x))}{\psi(x)}$$

but

$$\frac{(\forall x \varphi(x))^k \quad (\forall x (\varphi(x) \rightarrow \psi(x)))^k}{\delta(x) \rightarrow \psi^k(x)}$$

and in general $\not\vdash (\delta(x) \rightarrow \psi^k) \leftrightarrow \psi^k$. The lemma is proved by induction on p . To account for formulas in the induction, we use the notion $k(\varphi)$ from Definition 3.2, which is tailored precisely to let the induction go through. \dashv

Remark 3.4. The proof translation leaves all the structure invariant. Thus, there is a provably total (in \mathbf{S}_2^1) function f such that, if p is a U, n -proof of φ , then p^k is a proof of φ^k , where p^k has the following properties. All axioms in p^k are $\leq f(n, k)$ and all formulas ψ in p^k have $\rho(\psi) \leq f(n, k)$.

⁴More efficient translations on proofs are also available. However they are less uniform.

There are various reasons to give, why we want the notion of interpretability to be provably transitive, that is, provably $S \triangleright U$ whenever both $S \triangleright T$ and $T \triangleright U$. The obvious way of proving this would be by composing (doing the one after the other) two interpretations. Thus, if we have $j : S \triangleright T$ and $k : T \triangleright U$ we would like to have $j \circ k : S \triangleright U$ where $j \circ k$ denotes a natural composition of translations.

If we try to perform a proof as depicted in Figure 1, at a certain point we would like to collect the S -proofs p_1, \dots, p_m of the j -translated T -axioms used in a proof of a k -translation of an axiom u of U , and take the maximum of all such proofs. But to see that such a maximum exists, we precisely need Σ_1 -collection.

However, it is desirable to also reason about interpretability in the absence of $B\Sigma_1$. A trick is needed to circumvent the problem of the unprovability of transitivity (and many other elementary desiderata).

One way to solve the problem is by switching to a notion of interpretability where the needed collection has been built in. This is the notion of smooth (axioms) interpretability as in Definition 3.5. In this paper we shall mean by interpretability, unless mentioned otherwise, always smooth interpretability. In the presence of $B\Sigma_1$ this notion will coincide with the earlier defined notion of interpretability, as Theorem 3.6 tells us.

Definition 3.5. We define the notions of axioms interpretability \triangleright_a , theorems interpretability \triangleright_t , smooth axioms interpretability \triangleright_{sa} and smooth theorems interpretability \triangleright_{st} .

$$\begin{aligned} j : U \triangleright_a V &:= \forall v \exists p (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(p, v^j)) \\ j : U \triangleright_t V &:= \forall \varphi \forall p \exists p' (\text{Proof}_V(p, \varphi) \rightarrow \text{Proof}_U(p', \varphi^j)) \\ j : U \triangleright_{sa} V &:= \forall x \exists y \forall v \leq x \exists p \leq y (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(p, v^j)) \\ j : U \triangleright_{st} V &:= \forall x \exists y \forall \varphi \leq x \forall p \leq x \exists p' \leq y (\text{Proof}_V(p, \varphi) \rightarrow \text{Proof}_U(p', \varphi^j)) \end{aligned}$$

It is now easy to see that \triangleright_a is indeed provably transitive over very weak base theories. For \triangleright_t this follows almost directly from the definition.

Theorem 3.6. In S_2^1 we have all the arrows as depicted in Figure 2.

Proof. We shall only comment on the arrows that are not completely trivial.

- $T \vdash j : U \triangleright_a V \rightarrow j : U \triangleright_{sa} V$, if $T \vdash B\Sigma_1$. So, reason in T and suppose $\forall v \exists p (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(p, v^j))$. If we fix some x , we get $\forall v \leq x \exists p (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(p, v^j))$. By $B\Sigma_1$ we get the required $\exists y \forall v \leq x \exists p \leq y (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(p, v^j))$. It is not clear if $T \vdash B\Sigma_1^-$, parameter-free collection, is a necessary condition.

- $S_2^1 \not\vdash j : U \triangleright_a V \rightarrow j : U \triangleright_t V$. A counter-example is given in [18].

In S_2^1 :

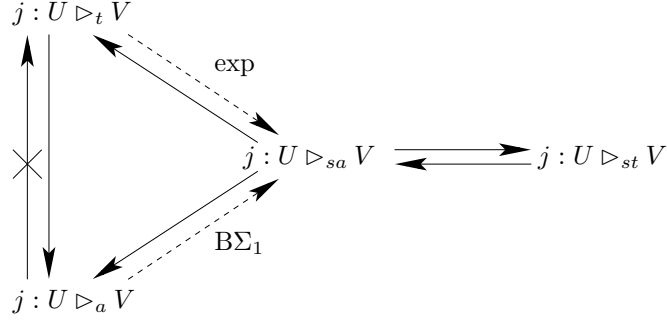


Figure 2: Versions of relative interpretability. The dotted arrows indicate that an additional condition is needed in our proof; the condition written next to it. The arrow with a cross through it, indicates that we know that the implication fails in S_2^1 .

• $T \vdash j : U \triangleright_t V \rightarrow j : U \triangleright_{sa} V$, if $T \vdash \text{exp}$. If V is reflexive, we get by Corollary 6.9 that $\vdash U \triangleright_t V \leftrightarrow U \triangleright_{sa} V$. However, different interpretations are used to witness the different notions of interpretability in this case. If $T \vdash \text{exp}$, we reason as follows. We reason in T and suppose that $\forall \varphi \forall p \exists p' (\text{Proof}_V(p, \varphi) \rightarrow \text{Proof}_U(p', \varphi^j))$. We wish to see

$$\forall x \exists y \forall v \leq x \exists p \leq y (\text{Axiom}_V(v) \rightarrow \text{Proof}_U(v^j)). \quad (1)$$

So, we pick x arbitrarily and consider⁵ $\nu := \bigwedge_{\text{Axiom}_V(v_i) \wedge v_i \leq x} v_i$. Notice that in the worst case, for all $y \leq x$, we have $\text{Axiom}_V(y)$, whence the length of ν can be bounded by $x \cdot |x|$. Thus, ν itself can be bounded by x^x , which exists whenever $T \vdash \text{exp}$. Clearly, $\exists p \text{Proof}_V(p, \nu)$ whence by our assumption $\exists p' \text{Proof}_U(p', \nu^j)$. In a uniform way, with just a slightly larger proof p'' , every v_i^j can be extracted from the proof p' of ν^j . We may take this $p'' \approx y$ to obtain (1). Note that $T \vdash \text{exp}$ is not a necessary condition since \triangleright_t implies \triangleright_a and if we have $B\Sigma_1$ the latter implies \triangleright_{sa} .

• $S_2^1 \vdash j : U \triangleright_{sa} V \rightarrow j : U \triangleright_{st} V$. So, we wish to see that

$$\forall x \exists y \forall \varphi \leq x \forall p \leq x \exists p' \leq y (\text{Proof}_V(p, \varphi) \rightarrow \text{Proof}_U(p', \varphi^j))$$

from the assumption that $j : U \triangleright_{sa} V$. So, we pick x arbitrarily. If now for some $p \leq x$ we have $\text{Proof}_V(p, \varphi)$, then clearly $\varphi \leq x$ and all axioms v_i of V that occur in p are $\leq x$. By our assumption $j : U \triangleright_{sa} V$, we can find a y_0 such that we can find proofs $p_i \leq y_0$ for all

⁵To see that ν exists, we seem to also use some collection; we collect all the $v_i \leq x$ for which $\text{Axiom}_V(v_i)$. However, it is not hard to see that we can consider ν also without collection since we use a natural coding.

the v_i^j . Now, with some sloppy notation, let $p^j[v_i^j/p_i]$ denote the j -translation of p where each j -translated axiom v_i^j is replaced by p_i .

Clearly, $p^j[v_i^j/p_i]$ is a proof for φ^j . The size of this proof can be estimated (again with sloppy notations):

$$p^j[v_i^j/p_i] \leq p^j[v_i^j/y_0] \leq (p^j)^{|y_0|} \leq (x^j)^{|y_0|}.$$

The latter bound is clearly present in S_2^1 . ⊥

We note that we have many admissible rules from one notion of interpretability to another. For example, by Buss's theorem on the provably total recursive functions of S_2^1 , it is not hard to see that

$$S_2^1 \vdash j : U \triangleright_a V \Rightarrow S_2^1 \vdash j : U \triangleright_t V.$$

In the rest of this paper, we shall at most places no longer write subscripts to the \triangleright 's. Our reading convention is then that we take that notion of interpretability that is best to perform the argument. Often this is just smooth interpretability \triangleright_s , which from now on is the notation for \triangleright_{sa} .

Moreover, in [18] some sort of conservation result concerning \triangleright_a and \triangleright_s is proved. For a considerable class of formulas φ and theories T , and for a considerable class of arguments we have that $T \vdash \varphi_a \Rightarrow T \vdash \varphi_s$. Here φ_a denotes the formula φ using the notion \triangleright_a and likewise for φ_s . Thus indeed, in many cases a sharp distinction between the notions involved is not needed.

We could also consider the following notion of interpretability.

$$j : U \triangleright_{st_1} V := \forall x \exists y \forall \varphi \leq x \exists p' \leq y (\Box_V \varphi \rightarrow \text{Proof}_U(p', \varphi^j))$$

Clearly, $j : U \triangleright_{st_1} V \rightarrow U \triangleright_{st} V$. However, for the reverse implication one seems to need $\text{B}\Pi_1^-$. Also, a straightforward proof of $U \vdash \text{id} : U \triangleright_{st_1} U$ seems to need $\text{B}\Pi_1^-$. Thus, the notion \triangleright_{st_1} seems to say more on the nature of a theory than on the nature of interpretability.

4 Cuts and induction

Inductive reasoning is a central feature of everyday mathematical practice. We are so used to it, that it enters a proof almost unnoticed. It is when one works with weak theories and in the absence of sufficient induction, that its all pervading nature is best felt.

A main tool to compensate for the lack of induction are the so-called definable cuts. They are definable initial segments of the natural numbers of a possibly non-standard model that possess some desirable properties that we could not infer for all numbers to hold by means of induction.

The idea is really simple. So, if we can derive $\varphi(0) \wedge \forall x (\varphi(x) \rightarrow \varphi(x+1))$ and do not have access to an induction axiom for φ , we just consider $J(x) : \forall y \leq x \varphi(y)$. Clearly J

now defines an initial segment on which φ holds. As we shall see, for a lot of reasoning we can restrict ourselves to initial segments rather than quantifying over all numbers.

4.1 Basic properties of cuts

Throughout the literature one can find some variations on the definition of a cut. At some places, a cut is only supposed to be an initial segment of the natural numbers. At other places some additional closure properties are demanded. By a well known technique due to Solovay (see for example [6]) any definable initial segment can be shortened in a definable way, so that it has a lot of desirable closure properties. Therefore, and as we almost always need the closure properties, we include them in our definition.

Definition 4.1. A definable U -cut is a formula $J(x)$ with only x free, for which we have the following.

1. $U \vdash J(0) \wedge \forall x (J(x) \rightarrow J(x+1))$
2. $U \vdash J(x) \wedge y \leq x \rightarrow J(y)$
3. $U \vdash J(x) \wedge J(y) \rightarrow J(x+y) \wedge J(x \cdot y)$
4. $U \vdash J(x) \rightarrow J(\omega_1(x))$

We shall sometimes also write $x \in J$ instead of $J(x)$. A first fundamental insight about cuts is the principle of *outside big, inside small*. Although not every number x is in J , we can find for every x a proof p_x that witnesses $x \in J$.

Lemma 4.2. *Let T and U be reasonable arithmetical theories and let J be a U -cut. We have that*

$$T \vdash \forall x \Box_U J(x).$$

Actually, we can have the quantifier over all cuts within the theory T , that is

$$T \vdash \forall^{U\text{-Cut}} J \forall x \Box_U J(x).$$

Proof. Let us start by making the quantifier $\forall^{U\text{-Cut}} J$ a bit more precise. By $\forall^{U\text{-Cut}} J$ we shall mean $\forall J (\Box_U \text{Cut}(J) \rightarrow \dots)$. Here $\text{Cut}(J)$ is the definable function that sends the code of a formula χ with one free variable to the code of the formula that expresses that χ defines a cut.

For a number a , we start with the standard proof of $J(0)$. This proof is combined with $a-1$ many instantiations of the standard proof of $\forall x (J(x) \rightarrow J(x+1))$. In the case of weaker theories, we have to switch to efficient numerals to keep the bound of the proof within range. \dashv

Remark 4.3. The proof sketch actually tells us that (provably in S_2^1) for every U -cut J , there is an $n \in \omega$ such that $\forall x \Box_{U,n} J(x)$.

Lemma 4.4. *Cuts are provably closed under terms, that is*

$$T \vdash \forall^{U\text{-Cut}} J \forall^{\text{Term}} t \Box_U \forall \vec{x} \in J \ t(\vec{x}) \in J.$$

Proof. By an easy induction on terms, fixing some U -cut J . Prima facie this looks like a Σ_1 -induction but it is easy to see that the proofs have poly-time (in t) bounds, whence the induction is $\Delta_0(\omega_1)$. \dashv

As all U -cuts are closed under $\omega_1(x)$ and the smash function \sharp , simply relativizing all quantors to a cut is an example of an interpretation of S_2^1 in U . We shall always denote both the cut and the interpretation that it defines by the same letter.

4.2 Cuts and the Henkin construction

It is well known that we can perform the Henkin construction in a rather weak meta-theory. As the Henkin model has a uniform description, we can link it to interpretations. The following theorem makes this precise.

Theorem 4.5. *If $U \vdash \text{Con}(V)$, then $U \triangleright V$.*

Early treatments of this theorem were given in [20] and [7]. A first fully formalized version was given in [2]. A proof of Theorem 4.5 would closely follow the Henkin construction.

Thus, first the language of V is extended so that it contains a witness $c_{\exists x \varphi(x)}$ for every existential sentence $\exists x \varphi(x)$. Then we can extend V to a maximal consistent V' in the enriched language, containing all sentences of the form $\exists x \varphi(x) \rightarrow \varphi(c_{\exists x \varphi(x)})$. This V' can be seen as a term model with a corresponding truth predicate. Clearly, if $V \vdash \varphi$ then $\varphi \in V'$. It is not hard to see that V' is representable (close inspection yields a Δ_2 -representation) in U .

At first sight the argument uses quite some induction in extending V to V' . Miraculously enough, the whole argument can be adapted to S_2^1 . The trick consists in replacing the use of induction by employing definable cuts as is explained above. We get the following theorem.

Theorem 4.6. *For any numberizable theories U and V , we have that*

$$S_2^1 \vdash \Box_U \text{Con}(V) \rightarrow \exists k (k : U \triangleright V \ \& \ \forall \varphi \Box_U (\Box_V \varphi \rightarrow \varphi^k)).$$

Proof. A proof can be found in [18]. Actually something stronger is proved there. Namely, that for some standard number m we have

$$\forall \varphi \exists p \leq \omega_1^m(\varphi) \text{ Proof}_U(p, \Box_V \varphi \rightarrow \varphi^k).$$

\dashv

As cuts have nice closure properties, many arguments can be performed within that cut. The numbers in the cut will, so to say, play the role of the normal numbers. It turns out that the whole Henkin argument can be carried out using only the consistency on a cut.

We shall write $\Box_T^J \varphi$ for $\exists p \in J \text{ Proof}_T(p, \varphi)$. Thus, it is also clear what $\Diamond_T^J \varphi$ and $\text{Con}^J(V)$ mean.

Theorem 4.7. *We have Theorem 4.6 also in the following form.*

$$T \vdash \forall^{U-\text{Cut}} I \left[\Box_U \text{Con}^I(V) \rightarrow \exists k (k : U \triangleright V \ \& \ \forall \varphi \Box_U (\Box_V \varphi \rightarrow \varphi^k)) \right]$$

Proof. By close inspection of the proof of Theorem 4.6. All operations on hypothetical proofs p can be bounded by some $\omega_1^k(p)$, for some standard k . As I is closed under $\omega_1(x)$, all the bounds remain within I . \dashv

We conclude this subsection with two asides, closely related to the Henkin construction.

Lemma 4.8. *Let U contain S_2^1 . We have that $U \vdash \text{Con}(\text{Pred})$. Here, $\text{Con}(\text{Pred})$ is a natural formalization of the statement that predicate logic is consistent.*

Proof. By defining a simple (one-point) model within S_2^1 . \dashv

Remark 4.9. If U proves $L\Delta_2^0$, then it holds that $U \triangleright V$ iff V is interpretable in U by some interpretation that maps identity to identity.

Proof. Suppose $j : U \triangleright V$ with $j = \langle \delta, F \rangle$. We can define $j' := \langle \delta', F' \rangle$ with $\delta'(x) := \delta(x) \wedge \forall y < x (\delta(y) \rightarrow y \neq^j x)$. F' agrees with F on all symbols except that it maps identity to identity. By the minimal number principle we can prove $\forall x (\delta(x) \rightarrow \exists x' (x' =^j x) \wedge \delta'(x))$, and thus $\forall \vec{x} (\delta'(\vec{x}) \rightarrow (\varphi^j(\vec{x}) \leftrightarrow \varphi^{j'}(\vec{x})))$ for all formulae φ . \dashv

5 Pudlák's lemma

In this section we will state and prove what is known as Pudlák's lemma. Moreover, we shall prove a very useful consequence of this lemma. Roughly speaking, Pudlák's lemma tells us how interpretations bear on the models that they induce. Therefor, let us first see how interpretations and models are related.

5.1 Interpretations and models

We can view interpretations $j : U \triangleright V$ as a way of defining uniformly a model \mathcal{N} of V inside a model \mathcal{M} of U . Interpretations in foundational papers mostly bear the guise of a uniform model construction.

Definition 5.1. Let $j : U \triangleright V$ with $j = \langle \delta, F \rangle$. If $\mathcal{M} \models U$, we denote by \mathcal{M}^j the following model.

- $|\mathcal{M}^j| = \{x \in |\mathcal{M}| \mid \mathcal{M} \models \delta(x)\} / \equiv$, where $a \equiv b$ iff $\mathcal{M} \models a =^j b$.
- $\mathcal{M}^j \models R(\alpha_1, \dots, \alpha_n)$ iff $\mathcal{M} \models F(R)(a_1, \dots, a_n)$, for some $a_1 \in \alpha_1, \dots, a_n \in \alpha_n$.

The fact that $j : U \triangleright V$ is now reflected in the observation that, whenever $\mathcal{M} \models U$, then $\mathcal{M}^j \models V$.

On many occasions viewing interpretations as uniform model constructions provides the right heuristics.

5.2 Pudlák's isomorphic cut

Pudlák's lemma is central to many arguments in the field of interpretability logics. It provides a means to compare a model \mathcal{M} of U and its internally defined model \mathcal{M}^j of V if $j : U \triangleright V$. If U has full induction, this comparison is fairly easy.

Theorem 5.2. *Suppose $j : U \triangleright V$ and U has full induction. Let \mathcal{M} be a model of U . We have that $\mathcal{M} \preceq_{\text{end}} \mathcal{M}^j$ via a definable embedding.*

Proof. If U has full induction and $j : U \triangleright V$, we may by Remark 4.9 actually assume that j maps identity in V to identity in U . Thus, we can define the following function.

$$f := \begin{cases} 0 \mapsto 0^j \\ x + 1 \mapsto f(x) +^j 1^j \end{cases}$$

Now, by induction, f can be proved to be total. Note that full induction is needed here, as we have a-priori no bound on the complexity of 0^j and $+^j$. Moreover, it can be proved that $f(a + b) = f(a) +^j f(b)$, $f(a \cdot b) = f(a) \cdot^j f(b)$ and that $y \leq^j f(b) \rightarrow \exists a < b \ f(a) = y$. In other words, that f is an isomorphism between its domain and its co-domain and the co-domain is an initial segment of \mathcal{M}^j . \dashv

If U does not have full induction, a comparison between \mathcal{M} and \mathcal{M}^j is given by Pudlák's lemma, first explicitly mentioned in [15]. Roughly, Pudlák's lemma says that in the general case, we can find a definable U -cut I of \mathcal{M} and a definable embedding $f : I \rightarrow \mathcal{M}^j$ such that $f[I] \preceq_{\text{end}} \mathcal{M}^j$.

In formulating the statement we have to be careful as we can no longer assume that identity is mapped to identity. A precise formulation of Pudlák's lemma in terms of an isomorphism between two initial segments can for example be found in [10]. We have chosen here to formulate and prove the most general syntactic consequence of Pudlák's lemma, namely that I and $f[I]$, as substructures of \mathcal{M} and \mathcal{M}^j respectively, make true the same Δ_0 -formulas.

In the proof of Pudlák's lemma we shall make the quantifier $\exists^{j, J\text{-function}} h$ explicit. It basically means that h defines a function from a cut J to the $=^j$ -equivalence classes of the numbers defined by the interpretation j .

Lemma 5.3 (Pudlák's Lemma).

$$S_2^1 \vdash j : U \triangleright V \rightarrow \exists^{U\text{-Cut}} J \exists^{j, J\text{-function}} h \forall^{\Delta_0} \varphi \square_U \forall \vec{x} \in J (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x}))$$

Moreover, the h and J can be obtained uniformly from j by a function that is provably total in S_2^1 .

Proof. Again, by $\exists^{U\text{-Cut}} J \psi$ we shall mean $\exists J (\square_U \text{Cut}(J) \wedge \psi)$, where $\text{Cut}(J)$ is the definable function that sends the code of a formula χ to the code of a formula that expresses that χ defines a cut. We apply a similar strategy for quantifying over j , J -functions. Given a translation j , the defining property for a relation H to be a j , J -function is

$$\forall \vec{x}, y, y' \in J (H(\vec{x}, y) \ \& \ H(\vec{x}, y') \rightarrow y =^j y').$$

We will often consider H as a function h and write for example $\psi(h(\vec{x}))$ instead of

$$\forall y (H(\vec{x}, y) \rightarrow \psi(y)).$$

The idea of the proof is very easy. Just map the numbers of U via h to the numbers of V so that 0 goes to 0^j and the mapping commutes with the successor relation. If we want to prove a property of this mapping, we might run into problems as the intuitive proof appeals to induction. And sufficient induction is precisely what we lack in weaker theories.

The way out here is to just put all the properties that we need our function h to possess into its definition. Of course, then the work is in checking that we still have a good definition. Being good means here that the set of numbers on which h is defined induces a definable U -cut.

In a sense, we want an (definable) initial part of the numbers of U to be isomorphic under h to an initial part of the numbers of V . Thus, h should definitely commute with successor, addition and multiplication. Moreover, the image of h should define an initial segment, that is, be closed under the smaller than relation. All these requirements are reflected in the definition of **Goodsequence**. Let δ denote the domain specifier of the translation j . We define

$$\begin{aligned} \text{Goodsequence}(\sigma, x, y) \quad := \quad & \text{lh}(\sigma) = x + 1 \wedge \sigma_0 =^j 0^j \wedge \sigma_x =^j y \\ & \wedge \forall i \leq x \ \delta(\sigma_i) \\ & \wedge \forall i < x \ (\sigma_{i+1} =^j \sigma_i +^j 1^j) \\ & \wedge \forall k + l \leq x \ (\sigma_k +^j \sigma_l =^j \sigma_{k+l}) \\ & \wedge \forall k \cdot l \leq x \ (\sigma_k \cdot^j \sigma_l =^j \sigma_{k \cdot l}) \\ & \wedge \forall a \ (a \leq^j y \rightarrow \exists i \leq x \ \sigma_i =^j a). \end{aligned}$$

Subsequently, we define

$$\begin{aligned} H(x, y) \quad := \quad & \exists \sigma \ \text{Goodsequence}(\sigma, x, y) \\ & \wedge \forall \sigma' \forall y' \ (\text{Goodsequence}(\sigma', x, y') \rightarrow y =^j y'), \end{aligned}$$

and

$$J'(x) := \forall x' \leq x \exists y H(x', y).$$

Finally, we define J to be the closure of J' under $+$, \cdot and $\omega_1(x)$. Now that we have defined all the machinery we can start the real proof. The reader is encouraged to see at what place which defining property is used in the proof.

We first note that $J'(x)$ indeed defines a U -cut. For $\Box_U J'(0)$ you basically need sequentiality of U , and the translations of the identity axioms and properties of 0.

To see $\Box_U \forall x (J'(x) \rightarrow J'(x+1))$ is also not hard. It follows from the translation of basic properties provable in V , like $x = y \rightarrow x+1 = y+1$ and $x + (y+1) = (x+y) + 1$, etc. The other properties of Definition 4.1 go similarly.

We should now see that h is a j, J -function. This is quite easy, as we have all the necessary conditions present in our definition. Actually, we have

$$\Box_U \forall x, y \in J (h(x) =^j h(y) \leftrightarrow x = y) \quad (2)$$

The \leftarrow direction reflects that h is a j, J -function. The \rightarrow direction follows from elementary reasoning in U using the translation of basic arithmetical facts provable in V . So, if $x \neq y$, say $x < y$, then $x + (z+1) = y$ whence $h(x) +^j h(z+1) =^j h(y)$ which implies $h(x) \neq^j h(y)$.

We are now to see that for our U -cut J and for our j, J -function h we indeed have that⁶

$$\forall^{\Delta_0} \varphi \Box_U \forall \vec{x} \in J (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})).$$

First we shall proof this using a seemingly Σ_1 -induction. A closer inspection of the proof shall show that we can provide at all places sufficiently small bounds, so that actually an $\omega_1(x)$ -induction suffices. We first proof the following claim.

Claim 1. $\forall^{\text{Term}} t \Box_U \forall \vec{x}, y \in J (t^j(h(\vec{x})) =^j h(y) \leftrightarrow t(\vec{x}) = y)$

Proof. The proof is by induction on t . The basis is trivial. To see for example

$$\Box_U \forall y \in J (0^j =^j h(y) \leftrightarrow 0 = y)$$

we reason in U as follows. By the definition of h , we have that $h(0) =^j 0^j$, and by (2) we moreover see that $0^j =^j h(y) \leftrightarrow 0 = y$. The other base case, that is, when t is an atom, is precisely (2).

For the induction step, we shall only do $+$, as \cdot goes almost completely the same. Thus, we assume that $t(\vec{x}) = t_1(\vec{x}) + t_2(\vec{x})$ and set out to prove

$$\Box_U \forall \vec{x}, y \in J (t_1^j(h(\vec{x})) +^j t_2^j(h(\vec{x})) =^j h(y) \leftrightarrow t_1(\vec{x}) + t_2(\vec{x}) = y).$$

Within U :

⁶We use $h(\vec{x})$ as short for $h(x_0), \dots, h(x_n)$.

← If $t_1(\vec{x}) + t_2(\vec{x}) = y$, then by Lemma 4.4, we can find y_1 and y_2 with $t_1(\vec{x}) = y_1$ and $t_2(\vec{x}) = y_2$. The induction hypothesis tells us that $t_1^j(h(\vec{x})) =^j h(y_1)$ and $t_2^j(h(\vec{x})) =^j h(y_2)$. Now by (2), $h(y_1 + y_2) =^j h(y)$ and by the definition of h we get that

$$\begin{aligned} h(y_1 + y_2) &=^j h(y_1) +^j h(y_2) \\ &=^j_{\text{i.h.}} t_1^j(h(\vec{x})) +^j t_2^j(h(\vec{x})) \\ &=^j (t_1(h(\vec{x})) + t_2(h(\vec{x})))^j. \end{aligned}$$

→ Suppose now $t_1^j(h(\vec{x})) +^j t_2^j(h(\vec{x})) =^j h(y)$. Then clearly $t_1^j(h(\vec{x})) \leq^j h(y)$ whence by the definition of h we can find some $y_1 \leq y$ such that $t_1^j(h(\vec{x})) =^j h(y_1)$ and likewise for t_2 (using the translation of the commutativity of addition). The induction hypothesis now yields $t_1(\vec{x}) = y_1$ and $t_2(\vec{x}) = y_2$. By the definition of h , we get $h(y) =^j h(y_1) +^j h(y_2) =^j h(y_1 + y_2)$, whence by (2), $y_1 + y_2 = y$, that is, $t_1(\vec{x}) + t_2(\vec{x}) = y$.

⊥

We now prove by induction on $\varphi \in \Delta_0$ that

$$\Box_U \forall \vec{x} \in J (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})). \quad (3)$$

For the base case, we consider that $\varphi \equiv t_1(\vec{x}) + t_2(\vec{x})$. We can now use Lemma 4.4 to note that

$$\Box_U \forall \vec{x} \in J (t_1(\vec{x}) = t_2(\vec{x}) \leftrightarrow \exists y \in J (t_1(\vec{x}) = y \wedge t_2(\vec{x}) = y))$$

and then use Claim 1, the transitivity of $=$ and its translation to obtain the result.

The boolean connectives are really trivial, so we only need to consider bounded quantification. We show (still within U) that

$$\forall y, \vec{z} \in J (\forall x \leq^j h(y) \varphi^j(x, h(\vec{z})) \leftrightarrow \forall x \leq y \varphi(x, \vec{z})).$$

← Assume $\forall x \leq y \varphi(x, \vec{z})$ for some $y, \vec{z} \in J$. We are to show $\forall x \leq^j h(y) \varphi^j(x, h(\vec{z}))$. Now, pick some $x \leq^j h(y)$ (the translation of the universal quantifier actually gives us an additional $\delta(x)$ which we shall omit for the sake of readability). Now by the definition of h we find some $y' \leq y$ such that $h(y') = x$. As $y' \leq y$, by our assumption, $\varphi(y', \vec{z})$ whence by the induction hypothesis $\varphi^j(h(y'), h(\vec{z}))$, that is $\varphi^j(x, h(\vec{z}))$. As x was arbitrarily $\leq^j h(y)$, we are done.

→ Suppose $\forall x \leq^j h(y) \varphi^j(x, h(\vec{z}))$. We are to see that $\forall x \leq y \varphi(x, \vec{z})$. So, pick $x \leq y$ arbitrarily. Clearly $h(x) \leq^j h(y)$, whence, by our assumption $\varphi^j(h(x), h(\vec{z}))$ and by the induction hypothesis, $\varphi(x, \vec{z})$.

Note that in our proof we have used twice a Σ_1 -induction; In Claim 1 and in proving (3). Let us now see that we can dispense with the Σ_1 induction.

In both cases, at every induction step, a constant piece p' of proof is added to the total proof. This piece looks every time the same. Only some parameters in it have to

be replaced by subterms of t . So, the addition to the total proof can be estimated by $p'_a(t)$ which is about $\mathcal{O}(t^k)$ for some standard k and indeed, our induction was really but a bounded one. Both our inductions went over syntax and whence are available in S_2^1 .

Note that in proving (3) we dealt with the bounded quantification by appealing to the induction hypothesis only once, followed by a generalization. So, fortunately we did not need to apply the induction hypothesis to all $x \leq y$, which would have yielded an exponential blow-up. \dashv

Remark 5.4. Pudlák's lemma is valid already if we employ the notion of theorems interpretability rather than smooth interpretability. If we work with theories in the language of arithmetic, we can do even better. In this case, axioms interpretability can suffice. In order to get this, all arithmetical facts whose translations were used in the proof of Lemma 5.3 have to be promoted to the status of axiom. However, a close inspection of the proof shows that these facts are very basic and that there are not so many of them.

If j is an interpretation with $j : \alpha \triangleright \beta$, we shall sometimes call the corresponding isomorphic cut that is given by Lemma 5.3, the *Pudlák cut* of j and denote it by the corresponding upper case letter J .

5.3 A consequence of Pudlák's Lemma

The following consequence of Pudlák's Lemma is simple, yet can be very useful. For simplicity we state the consequence for sentential extensions of some base theory T extending S_2^1 . Thus, $\alpha \triangleright \beta$ will be short for $(T + \alpha) \triangleright (T + \beta)$.

Lemma 5.5. (In S_2^1 .) *If $j : \alpha \triangleright \beta$ then, for every $T + \beta$ cut I there exists a $T + \alpha$ cut J such that for every γ we have that*

$$j : \alpha \wedge \Box^J \gamma \triangleright \beta \wedge \Box^I \gamma.$$

Proof. By a minor adaptation of the standard argument. First, we define **Goodsequence**.

$$\begin{aligned} \text{Goodsequence}(\sigma, x, y) \quad &:= \quad \text{lh}(\sigma) = x + 1 \wedge \sigma_0 = {}^j 0^j \wedge \sigma_x = {}^j y \\ &\wedge \forall i < x \quad (\sigma_{i+1} = {}^j \sigma_i + {}^j 1^j) \\ &\wedge \forall k + l \leq x \quad (\sigma_k + {}^j \sigma_l = {}^j \sigma_{k+l}) \\ &\wedge \forall k \cdot l \leq x \quad (\sigma_k \cdot {}^j \sigma_l = {}^j \sigma_{k \cdot l}) \\ &\wedge \forall a \quad (a \leq {}^j y \rightarrow \exists i \leq x \quad \sigma_i = {}^j a) \\ &\wedge \forall i < x \quad I^j(\sigma_i) \end{aligned}$$

Next, we define

$$\begin{aligned} H(x, y) \quad &:= \quad \exists \sigma \text{ Goodsequence}(\sigma, x, y) \\ &\wedge \forall \sigma' \forall y' \quad (\text{Goodsequence}(\sigma', x, y') \rightarrow y = {}^j y'), \end{aligned}$$

and

$$J'(x) := \forall x' \leq x \exists y H(x', y).$$

Finally, we define J to be the closure of J' under $+$, \cdot and $\omega_1(x)$.

As before, one can see $H(x, y)$ as defining a function (modulo $=^j$), call it h , that defines an isomorphism between J and the image of J . Moreover, in the definition of **Goodsequence** we demanded that the image of h is a subset of I in the clause $\forall i < x I^j(\sigma_i)$.

It is easy to see that J' is closed under successor, that is, $J'(x) \rightarrow J'(x+1)$. We only comment on the new ingredient of the image of h being a subset of I . However, $T + \beta \vdash I(x) \rightarrow I(x+1)$, as I is a definable cut. As $j : T + \alpha \triangleright T + \beta$, clearly $T + \alpha \vdash I^j(x) \rightarrow I^j(x +^j 1^j)$ and indeed J' is closed under successor.

⊥

In the literature, Lemma 5.5 was known only for I to be the trivial cut of all numbers defined by $x = x$.

6 The Orey-Hájek characterizations

This final section contains the most substantial part of the paper. We consider the diagram from Figure 3. It is well known that all the implications hold when both U and V are reflexive. This fact is referred to as the Orey-Hájek characterizations ([2], [14], [4], [5]) for interpretability. However, for the Π_1 -conservativity part, we should also mention work by Guaspari, Lindström and Pudlák ([3], [12], [13], [15]).

In this section we shall comment on all the implications in Figure 3, and study the conditions on U , V and the meta-theory, that are necessary or sufficient.

Lemma 6.1. *In S_2^1 we can prove $\forall n \square_U \text{Con}_n(V) \rightarrow U \triangleright V$.*

Proof. The only requirement for this implication to hold, is that $U \vdash \text{Con}(\text{Pred})$. But, by our assumptions on U and by Lemma 4.8 this is automatically satisfied.

Let us first give the informal proof. Thus, let $\text{Axiom}_V(x)$ be the formula that defines the axiom set of V .

We now apply a trick due to Feferman and consider the theory V' that consists of those axioms of V up to which we have evidence for their consistency. Thus, $\text{Axiom}_{V'}(x) := \text{Axiom}_V(x) \wedge \text{Con}_x(V)$.

We shall now prove that $U \triangleright V$ in two steps. First, we will see that

$$U \vdash \text{Con}(V'). \tag{4}$$

Thus, by Theorem 4.5 we get that $U \triangleright V'$. Second, we shall see that

$$V = V'. \tag{5}$$

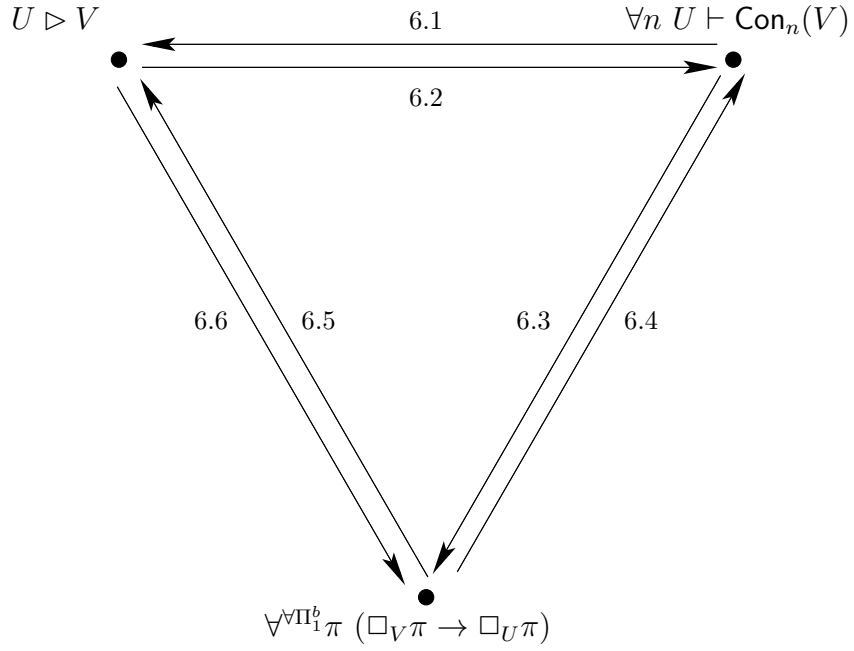


Figure 3: Characterizations of interpretability. The labels at the arrows are references to where in the paper this arrow is proven and what the conditions are for the arrow to hold. Moreover, we will discuss which conditions should hold for the base theory so that the implications become formalizable.

To see (4), we reason in U , and assume for a contradiction that $\text{Proof}_{V'}(p, \perp)$ for some proof p . We consider the largest axiom v that occurs in p . By assumption we have (in U) that $\text{Axiom}_{V'}(v)$ whence $\text{Con}_v(V)$. But, as clearly $V' \subseteq V$, we see that p is also a V -proof. We can now obtain a cut-free proof p' of \perp . Clearly $\text{Proof}_{V,v}(p', \perp)$ and we have our contradiction.

If V' is empty, we cannot consider v . But in this case, $\text{Con}(V') \leftrightarrow \text{Con}(\text{Pred})$, and by assumption, $U \vdash \text{Con}(\text{Pred})$.

We shall now see (5). Clearly $\mathbb{N} \models \text{Axiom}_{V'}(v) \rightarrow \text{Axiom}_V(v)$ for any $v \in \mathbb{N}$. To see that the converse also holds, we reason as follows.

Suppose $\mathbb{N} \models \text{Axiom}_V(v)$. By assumption $U \vdash \text{Con}_v(V)$, whence $\text{Con}_v(V)$ holds on any model \mathcal{M} of U . We now observe that \mathbb{N} is an initial segment of (the numbers of) any model \mathcal{M} of U , that is,

$$\mathbb{N} \preceq_{\text{end}} \mathcal{M}. \quad (6)$$

As $\mathcal{M} \models \text{Con}_v(V)$ and as $\text{Con}_v(V)$ is a Π_1 -sentence, we see that also $\mathbb{N} \models \text{Con}_v(V)$. By assumption we had $\mathbb{N} \models \text{Axiom}_V(v)$, thus we get that $\mathbb{N} \models \text{Axiom}_{V'}(v)$. We conclude that

$$\mathbb{N} \models \text{Axiom}_V(x) \leftrightarrow \text{Axiom}_{V'}(x) \quad (7)$$

whence, that $V = V'$. As $U \vdash \text{Con}(V')$, we get by Theorem 4.5 that $U \triangleright V'$. We may thus infer the required $U \triangleright V$.

It is not possible to directly formalize the informal proof. At (7) we concluded that $V = V'$. This actually uses some form of Π_1 -reflection which is manifested in (6). The lack of reflection in the formal environment will be compensated by another sort of reflection, as formulated in Theorem 4.6.

Moreover, to see (4), we had to use a cut elimination. To avoid this, we shall need a sharper version of Feferman's trick.

Let us now start with the formal proof sketch and refer to [18] for more details. We shall reason in U . Without any induction we conclude $\forall x (\text{Con}_x(V) \rightarrow \text{Con}_{x+1}(V))$ or $\exists x (\text{Con}_x(V) \wedge \Box_{V,x+1} \perp)$. In both cases we shall sketch a Henkin construction.

If $\forall x (\text{Con}_x(V) \rightarrow \text{Con}_{x+1}(V))$ and also $\text{Con}_0(V)$, we can find a cut $J(x)$ with $J(x) \rightarrow \text{Con}_x(V)$. We now consider the following non-standard proof predicate.

$$\Box_W^* \varphi := \exists x \in J \Box_{W,x} \varphi$$

We note that we have $\text{Con}^*(V)$, where $\text{Con}^*(V)$ of course denotes $\neg(\exists x \in J \Box_{V,x} \perp)$. As always, we extend the language on J by adding witnesses and define a series of theories in the usual way. That is, by adding more and more sentences (in J) to our theories while staying consistent (in our non-standard sense).

$$V = V_0 \subseteq V_1 \subseteq V_2 \subseteq \dots \text{ with } \text{Con}^*(V_i) \quad (8)$$

We note that $\Box_{V_i}^* \varphi$ and $\Box_{V_i}^* \neg \varphi$ is not possible, and that for $\varphi \in J$ we can not have $\text{Con}^*(\varphi \wedge \neg \varphi)$. These observations seem to be too trivial to make, but actually many a non-standard proof predicate encountered in the literature does prove the consistency of inconsistent theories.

As always, the sequence (8) defines a cut $I \subseteq J$, that induces a Henkin set W and we can relate our required interpretation k to this Henkin set as was, for example, done in [18].

We now consider the case that for some fixed b we have $\text{Con}_b(V) \wedge \Box_{V,b+1} \perp$. We note that we can see the uniqueness of this b without using any substantial induction. Basically, we shall now do the same construction as before only that we now possibly stop at b .

For example the cut $J(x)$ will now be replaced by $x \leq b$. Thus, we may end up with a truncated Henkin set W . But this set is complete with respect to relatively small formulas. Moreover, W is certainly closed under subformulas and substitution of witnesses. Thus, W is sufficiently large to define the required interpretation k .

In both cases we can perform the following reasoning.

$$\begin{aligned}
\Box_V \varphi &\rightarrow \exists x \Box_{V,x} \varphi \\
&\rightarrow \exists x \Box_U (\text{Con}_x(V) \wedge \Box_{V,x} \varphi) \\
&\rightarrow \Box_U \Box_V^* \varphi \\
&\rightarrow \Box_U \varphi^k && \text{by Theorem 4.6.}
\end{aligned}$$

The remarks from [18] on the bounds of our proofs are still applicable and we thus obtain a smooth interpretation. \dashv

Lemma 6.2. *In the presence of exp , we can prove that for reflexive U , $U \triangleright V \rightarrow \forall x \Box_U \text{Con}_x(V)$.*

Proof. The informal argument is conceptually very clear and we have depicted it in Figure 4. The accompanying reasoning is as follows.

We assume $U \triangleright V$, whence for some k we have $k : U \triangleright V$. Thus, for axioms interpretability we find that $\forall u \exists p (\text{Axiom}_V(u) \rightarrow \text{Proof}_U(p, u^k))$. We are now to see that $\forall x U \vdash \text{Con}_x(V)$. So, we fix some x . By our assumption we get that for some l , that

$$\forall u \leq x \exists p (\text{Axiom}_V(u) \rightarrow \text{Proof}_{U,l}(p, u^k)). \quad (9)$$

This formula is actually equivalent to the Σ_1 -formula

$$\exists n \forall u \leq x \exists p \leq n (\text{Axiom}_V(u) \rightarrow \text{Proof}_{U,l}(p, u^k)) \quad (10)$$

from which we may conclude by provable Σ_1 -completeness,

$$U \vdash \exists n \forall u \leq x \exists p \leq n (\text{Axiom}_V(u) \rightarrow \text{Proof}_{U,l}(p, u^k)). \quad (11)$$

In U :

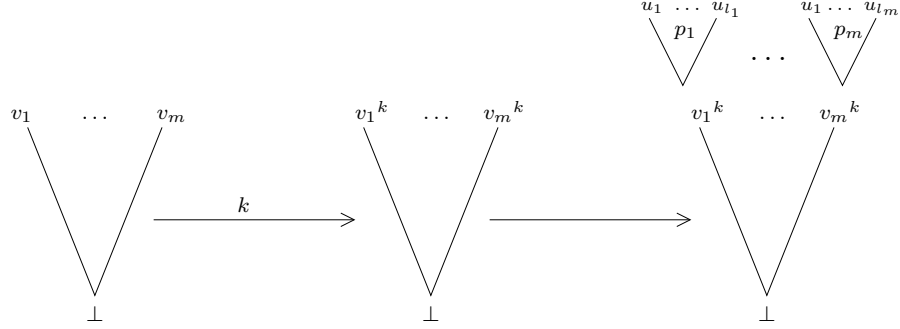


Figure 4: Transformations on proofs

We now reason in U and suppose that there is some V, x -proof p of \perp . The assumptions in p are axioms $v_1 \dots v_m$ of V , with each $v_i \leq x$. Moreover, all the formulas ψ in p have $\rho(\psi) \leq x$. By Lemma 3.3, this p transforms to a proof p^k of \perp^k which is again \perp .

The assumptions in p^k are now among the $v_1^k \dots v_m^k$. By Remark 3.4 we get that for some n' depending on x and k , we have that all the axioms in p^k are $\leq n'$ and all the ψ occurring in p^k have $\rho(\psi) \leq n'$.

Now by (11), we have U, l -proofs $p_i \leq n$ of v_i^k . The assumptions in the p_i are axioms of U . Clearly all of these axioms are $\leq l$. We can now form a $U, l+n'$ -proof p' of \perp by substituting all the p_i for the $(v_i)^k$. Thus we have shown $\text{Proof}_{U, l+n'}(p', \perp)$. But this clearly contradicts the reflexivity of U .

The informal argument is readily formalized to obtain $T \vdash U \triangleright V \rightarrow \forall x \Box_U \text{Con}(V, x)$. However there are some subtleties.

First of all, to conclude that (9) is equivalent to (10), a genuine application of $\text{B}\Sigma_1$ is needed. If U lacks $\text{B}\Sigma_1$, we have to switch to smooth interpretability to still have the implication valid. Smoothness then automatically also provides the l that we used in 9.

In addition we need that T proves the totality of exponentiation. For weaker theories, we only have provable $\exists \Sigma_1^b$ -completeness. But if $\text{Axiom}_V(u)$ is Δ_1^b , we can only guarantee that $\forall u \leq m \exists p \leq n (\text{Axiom}_V(u) \rightarrow \text{Proof}_U(p, u^k))$ is Π_2^b . As far as we know, exponentiation is needed to prove $\exists \Pi_2^b$ -completeness.

All other transformations of objects in our proof only require the totality of $\omega_1(x)$. \dashv

The assumption that U is reflexive can in a sense not be dispensed with. That is, if

$$\forall V (U \triangleright V \rightarrow \forall x \Box_U \text{Con}_x(V)), \quad (12)$$

then U is reflexive, as clearly $U \triangleright U$. In a similar way we see that if

$$\forall U (U \triangleright V \rightarrow \forall x \Box_U \text{Con}_x(V)), \quad (13)$$

then V is reflexive. However, V being reflexive could never be a sufficient condition for (13) to hold, as we know from [16] that interpreting reflexive theories in finitely many axioms is complete Σ_3 .

Lemma 6.3. *In S_2^1 we can prove $\forall x \Box_U \text{Con}_x(V) \rightarrow \forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi)$.*

Proof. There are no conditions on U and V for this implication to hold. We shall directly give the formal proof as the informal proof does not give a clearer picture.

Thus, we reason in S_2^1 and assume $\forall x \Box_U \text{Con}_x(V)$. Now we consider any $\pi \in \forall\Pi_1^b$ such that $\Box_V \pi$. Thus, for some x we have $\Box_{V,x} \pi$. We choose x large enough, so that we also have (see Remark 2.3)

$$\Box_U (\neg \pi \rightarrow \Box_{V,x} \neg \pi). \quad (14)$$

As $\Box_{V,x} \pi \rightarrow \Box_U \Box_{V,x} \pi$, we also have that

$$\Box_U \Box_{V,x} \pi. \quad (15)$$

Combining (14), (15) and the assumption that $\forall x \Box_U \text{Con}_x(V)$, we see that indeed $\Box_U \pi$. \dashv

Lemma 6.4. *In S_2^1 we can prove that for reflexive V we have*

$$\forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi) \rightarrow \forall x \Box_U \text{Con}_x(V).$$

Proof. If V is reflexive and $\forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi)$ then, as for every x , $\text{Con}_x(V)$ is a $\forall\Pi_1^b$ -formula, also $\forall x \Box_U \text{Con}_x(V)$. \dashv

It is obvious that

$$\forall U [\forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi) \rightarrow \forall x \Box_U \text{Con}_x(V)] \quad (16)$$

implies that V is reflexive. Likewise,

$$\forall V [\forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi) \rightarrow \forall x \Box_U \text{Con}_x(V)] \quad (17)$$

implies that U is reflexive. However, U being reflexive can never be a sufficient condition for (17) to hold. An easy counterexample is obtained by taking U to be PRA and V to be IS_1 as it is well-known that IS_1 is provably Π_2 conservative over PRA and that IS_1 is finitely axiomatized.

Lemma 6.5. *(In S_2^1 ;) For reflexive V we have $\forall^{\forall\Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi) \rightarrow U \triangleright V$.*

Proof. We know of no direct proof of this implication. Also, all proofs in the literature go via Lemmata 6.4 and 6.1, and hence use reflexivity of V . \dashv

In our context, the reflexivity of V is not necessary, as $\forall U \ U \triangleright S_2^1$ and S_2^1 is not reflexive.

Lemma 6.6. *Let U be a reflexive and sequential theory. We have in S_2^1 that $U \triangleright V \rightarrow \forall^{\forall \Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi)$.*

If moreover $U \vdash \text{exp}$ we also get $U \triangleright V \rightarrow \forall^{\Pi_1} \pi (\Box_V \pi \rightarrow \Box_U \pi)$. If U is not reflexive, we still have that $U \triangleright V \rightarrow \exists^{U\text{-Cut}} J \forall^{\Pi_1} \pi (\Box_V \pi \rightarrow \Box_U \pi^J)$.

For these implications, it is actually sufficient to work with the notion of theorems interpretability.

Proof. The intuition for the formal proof comes from Pudlák's lemma, which in turn is tailored to compensate a lack of induction. We shall first give an informal proof sketch if U has full induction. Then we shall give the formal proof using Pudlák's lemma.

If U has full induction and $j : U \triangleright V$, we may assume by Remark 4.9 assume that j maps identity to identity. Let \mathcal{M} be an arbitrary model of U . By Theorem 5.2 we now see that $\mathcal{M} \preceq_{\text{end}} \mathcal{M}^j$. If for some $\pi \in \Pi_1$, $\Box_V \pi$ then by soundness $\mathcal{M}^j \models \pi$, whence $\mathcal{M} \models \pi$. As \mathcal{M} was an arbitrary model of U , we get by the completeness theorem that $\Box_U \pi$.

To transform this argument into a formal one, valid for weak theories, there are two major adaptations to be made. First, the use of the soundness and completeness theorem have to be avoided. This can be done by simply staying in the realm of provability. Secondly, we should get rid of the use of full induction. This is done by switching to a cut in Pudlák's lemma.

Thus, the formal argument runs as follows. Reason in S_2^1 and assume $U \triangleright V$.

We fix some $j : U \triangleright V$. By Pudlák's lemma, Lemma 5.3, we now find⁷ a definable U -cut J and a j, J -function h such that

$$\forall^{\Delta_0} \varphi \Box_U \forall \vec{x} \in J (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})).$$

We shall see that for this cut J we have that

$$\forall^{\Pi_1} \pi (\Box_V \pi \rightarrow \Box_U \pi^J). \quad (18)$$

Therefore, we fix some $\pi \in \Pi_1$ and assume $\Box_V \pi$. Let $\varphi(x) \in \Delta_0$ be such that $\pi = \forall x \varphi(x)$. Thus we have $\Box_V \forall x \varphi(x)$, hence by theorems interpretability

$$\Box_U \forall x (\delta(x) \rightarrow \varphi^j(x)). \quad (19)$$

We are to see

$$\Box_U \forall x (J(x) \rightarrow \varphi(x)). \quad (20)$$

⁷Remark 5.4 ensures us that we can find them also in the case of theorems interpretability.

To see this, we reason in U and fix x such that $J(x)$. By definition of J , $h(x)$ is defined. By the definition of h , we have $\delta(h(x))$, whence by (19), $\varphi^j(h(x))$. Pudlák's lemma now yields the desired $\varphi(x)$. As x was arbitrary, we have proved (20).

So far, we have not used the reflexivity of U . We shall now see that

$$\forall^{\forall\Pi_1^b}\pi (\Box_U\pi^J \rightarrow \Box_U\pi)$$

holds for any U -cut J whenever U is reflexive. For this purpose, we fix some $\pi \in \forall\Pi_1^b$, some U -cut J and assume $\Box_U\pi^J$. Thus, $\exists n \Box_{U,n}\pi^J$ and also $\exists n \Box_U\Box_{U,n}\pi^J$. If $\pi = \forall x \varphi(x)$ with $\varphi(x) \in \Pi_1^b$, we get $\exists n \Box_U\Box_{U,n}\forall x (x \in J \rightarrow \varphi(x))$, whence also

$$\exists n \Box_U\forall x \Box_{U,n}(x \in J \rightarrow \varphi(x)).$$

By Lemma 4.2 and Remark 4.3, for large enough n , this implies

$$\exists n \Box_U\forall x \Box_{U,n}\varphi(x)$$

and by Lemma 2.4 (only here we use that $\pi \in \forall\Pi_1^b$) we obtain the required $\Box_U\forall x \varphi(x)$. \dashv

Again, by [16] we note that V being reflexive can never be a sufficient condition for $\forall U [U \triangleright V \rightarrow \forall^{\forall\Pi_1^b}\pi (\Box_V\pi \rightarrow \Box_U\pi)]$.

The main work on the Orey-Hájek characterization has now been done. We can easily extract some useful, mostly well-known corollaries.

Corollary 6.7. *If U is a reflexive theory, then*

$$T \vdash U \triangleright V \leftrightarrow \forall x \Box_U \text{Con}_x(V).$$

Here T contains exp and \triangleright denotes smooth interpretability.

Corollary 6.8. *(In S_2^1 .) If V is a reflexive theory, then the following are equivalent.*

1. $U \triangleright V$
2. $\exists^{U\text{-Cut}} J \forall^{\Pi_1}\pi (\Box_V\pi \rightarrow \Box_U\pi^J)$
3. $\exists^{U\text{-Cut}} J \forall x \Box_U \text{Con}_x^J(V)$

Proof. This is part of Theorem 2.3 from [16]. (1) \Rightarrow (2) is already proved in Lemma 6.6, (2) \Rightarrow (3) follows from the transitivity of V and (3) \Rightarrow (1) is a sharpening of Lemma 6.1. which closely follows Theorem 4.7. Note that \triangleright may denote smooth or theorems interpretability. \dashv

Corollary 6.9. *If V is reflexive, then*

$$S_2^1 \vdash U \triangleright_t V \leftrightarrow U \triangleright_s V.$$

Proof. By Remark 5.4 and Corollary 6.8. ⊢

Corollary 6.10. *If U and V are both reflexive theories we have that the following are provably equivalent in S_2^1 .*

1. $U \triangleright V$
2. $\forall^{\forall \Pi_1^b} \pi (\Box_V \pi \rightarrow \Box_U \pi)$
3. $\forall x \Box_U \text{Con}_x(V)$

Proof. If we go (1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1) we do not need the totality of **exp** that was needed for (1) \Rightarrow (3). ⊢

As an application we can, for example, see that $\text{PA} \triangleright \text{PA} + \text{InCon}(\text{PA})$. It is well known that PA is essentially reflexive which means that any finite extension of it is reflexive. So, we use Corollary 6.10 and, it is sufficient to show that $\text{PA} + \text{InCon}(\text{PA})$ is Π_1 -conservative over PA.

So, suppose that $\text{PA} + \text{InCon}(\text{PA}) \vdash \pi$ for some Π_1 -sentence π . In other words $\text{PA} \vdash \Box \perp \rightarrow \pi$. We shall now see that $\text{PA} \vdash \Box \pi \rightarrow \pi$, which by Löb's Theorem gives us $\text{PA} \vdash \pi$.

Thus, in PA, assume $\Box \pi$. Suppose for a contradiction that $\neg \pi$. By Σ_1 -completeness we also get $\Box \neg \pi$, which yields $\Box \perp$ with the assumption $\Box \pi$. But we have $\Box \perp \rightarrow \pi$ and we conclude π . A contradiction, so that indeed $\text{PA} \triangleright \text{PA} + \text{InCon}(\text{PA})$.

Acknowledgements

I am grateful to Lev Beklemishev, Félix Lara and Albert Visser for pointers to the literature and helpful discussions.

This research has been funded by Grant 2014 SGR 437 from the Catalan government and by Grant MTM2014-59178-P from the Spanish government.

References

- [1] S.R. Buss. First-order proof theory of arithmetic. In S.R. Buss, editor, *Handbook of Proof Theory*, pages 79–148, Amsterdam, 1998. Elsevier, North-Holland.
- [2] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.
- [3] D. Guaspari. Partially conservative sentences and interpretability. *Transactions of AMS*, 254:47–68, 1979.
- [4] P. Hájek. On interpretability in set theories I. *Comm. Math. Univ. Carolinae*, 12:73–79, 1971.

- [5] P. Hájek. On interpretability in set theories II. *Comm. Math. Univ. Carolinae*, 13:445–455, 1972.
- [6] P. Hájek and P. Pudlák. *Metamathematics of First Order Arithmetic*. Springer-Verlag, Berlin, Heidelberg, New York, 1993.
- [7] D. Hilbert and P. Bernays. *Grundlagen der Mathematik, Vols. I and II, 2d ed.* Springer-Verlag, Berlin, 1968.
- [8] J. J. Joosten. *Interpretability Formalized*. PhD thesis, Utrecht University, 2004.
- [9] J. J. Joosten. Two series of formalized interpretability principles for weak systems of arithmetic. *arXiv:1503.09130 [math.LO]*, 2015.
- [10] J.J. Joosten and A. Visser. The interpretability logic of *all* reasonable arithmetical theories. *Erkenntnis*, 53(1–2):3–26, 2000.
- [11] Jan Krajíček. *Bounded Arithmetic, Propositional Logic, and Complexity Theory*. Cambridge University Press, 1995.
- [12] P. Lindström. Some results on interpretability. In *Proceedings of the 5th Scandinavian Logic Symposium*, pages 329–361. Aalborg University press, 1979.
- [13] P. Lindström. On partially conservative sentences and interpretability. *Proceedings of the AMS*, 91(3):436–443, 1984.
- [14] S. Orey. Relative interpretations. *Zeitschrift f. math. Logik und Grundlagen d. Math.*, 7:146–153, 1961.
- [15] P. Pudlák. Cuts, consistency statements and interpretations. *Journal of Symbolic Logic*, 50:423–441, 1985.
- [16] V.Yu. Shavrukov. Interpreting reflexive theories in finitely many axioms. *Fundamenta Mathematicae*, 152:99–116, 1997.
- [17] A. Tarski, A. Mostowski, and R. Robinson. *Undecidable theories*. North-Holland, Amsterdam, 1953.
- [18] A. Visser. The formalization of interpretability. *Studia Logica*, 50(1):81–106, 1991.
- [19] A. Visser. The unprovability of small inconsistency. *Archive for Mathematical Logic*, 32:275–298, 1993.
- [20] H Wang. Arithmetical models of formal systems. *Methodos 3*, pages 217–232, 1951.